Big Data Analytics for Electricity Theft Detection in Smart Grids

Inam Ullah Khan[1,*], Nadeem Javaid[2], C. James Taylor[1], Kelum A. A. Gamage[3] and Xiandong Ma[1]
[1]*Engineering Department, Lancaster University, Bailrigg, Lancaster LA1 4YW, UK*
[2]*Department of Computer Science, COMSATS University Islamabad, Islamabad, 44000, Pakistan*
[3]*James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK*
*Correspondence: i.u.khan@lancaster.ac.uk, inamkhan@cuilahore.edu.pk

***Abstract*—In a smart grid, Electricity Theft Detection (ETD) is of great importance because it makes the smart grid cost-efficient. Existing methods for ETD cannot efficiently handle data imbalance, missing values, variance and non-linear data problems in smart meter data. Hence, an integrated strategy is needed to efficiently address underlying problems and accurately detect electricity theft using big data. In this work, a simple yet effective approach is adopted by integrating two different modules, namely data pre-processing and classification, in a single framework. Particularly, data prepressing module involves data imputation, outliers handling, data standardization, and class balancing steps to generate quality data for better classifier's training. The second module classifies honest and dishonest users with a Support Vector Machine (SVM) classifier. To improve the classifier's learning trend and accuracy, a Bayesian Optimizer is used to tune SVM's hyperparameters. Simulation results confirm that the proposed framework for ETD has superior performance in terms of accuracy than standard methods.**

***Index Terms*—Big data, Electricity theft detection, Feature engineering, Data classification, Smart grid.**

I. INTRODUCTION

In a deregulated environment of the power industry, the role of electricity theft detection (ETD) has become increasingly important in the smart grid. Electricity theft is one of the smart grid's main drivers, often causes a wide range of animalities at the planning and distribution level, and the advanced methods for ETD based on big data is always an essential and challenging issue. The primary purpose of ETD is to minimize non-technical losses and balance the supply-demand gap. An accurate ETD method brings an extraordinary level of flexibility for energy management and forms a win-win situation for generation and consumption [1], [2].

In a power network, customers have a predefined power purchase threshold and due to non-technical losses, the burden on end-users is ultimately increased. With ETD, utilities can control power demand for a specific time to get financial benefits in terms of energy generation cost savings. A precise ETD method reduces the demand-supply gap and helps develop a stable and efficient power management system. On one end, it helps the utility to address uncertain power generation challenges, specifically when penetration of RES is increasing. Besides, it brings higher reliability and aims to achieve available energy sources economically and rationally in an effective manner [3], [4].

Accurate ETD methods are of great importance for smart grids, and many intricate factors in big data would exacerbate the difficulty. The big data phenomenon is dynamic and complex that involve distinctive aspects of the time series data where the variation trends over time are non-linear. Accurate ETD is essential, but it is challenging to increase scalability, robustness, and accuracy due to the widespread non-linear data. Smart meters continuously monitor the associated factors such as time, consumption pattern, etc., all in real-time due to which the amount of data available for ETD is significantly big and hence challenging to handle, especially for ETD [5].

For ETD, authors in [6] proposed hybridization of Long Short-Term Memory (LSTM) and Multi-Layer Perceptron (MLP) techniques. The MLP is used for auxiliary data, whereas time sequence electricity data is evaluated with LSTM. The authors achieve good prediction results; however, model performance could further be improved if the class imbalance problem was solved during the data preparation stage. The model performance is relatively high on fewer data training in terms of False Positive Rate (FPR). However, when the data input for model training is high, its performance degraded and achieved only 54.5% performance metric for Precision-Recall Area Under the Curve (PR-AUC).

In [7], authors attempted to detect NTL in smart grids with Support Vector Machines (SVM) and a well-known boosting classifier named XGBoost. With smart meter data analysis, consumers are ranked based on load profiles. Afterward, essential features were extracted from auxiliary data. The SVM utilized empirical risk minimzation principal to imrove the training process. Afterward, the boosting algorithm utilized ensemble techniques to enhance classification performance. In the propose dstartegy, the authors did not take into account the pre-processing and data preparation steps. Like any other machine learning algorithm, SVM's performance could improve if refined data was fed into the classifier for training.

In [8], Shuan et al. used a well-known deep neural network model named Convolutional Neural Network (CNN) to detect electricity theft accurately. However, a major drawback of model generalization arises when classification output in CNN is taken through a fully connected layer. For this purpose, authors in [9] used a Random Forest (RF) classifier for final classification. In this model, the imbalanced class problem is solved with Synthetic Minority Oversampling Technique

978-1-5386-1370-2/18/$31.00 ©2018 IEEE

(SMOTE). CNN, along with RF, achieved better generalization; however, SMOTE's major drawback is synthetic data generation that pushes the model to overfit.

For ETD strategies, most of the current work is based on selection or classification approaches where Machine Learning (ML) and Artificial Neural Networks (ANNs) have shown improved performance. Nevertheless, both methods have limited abilities, such as ML models have low detection rate and high FPR, fail to deal with imbalance class, and face overfitting problems. When a model overfits, it means that model performance is good in training but not in classification. Similarly, ANN models have limited generalization capabilities, sensitivity to erroneous values, limited control over convergence/stability, and limited abilities to deal with the uncertainty. Besides, the learning-based model does not consider the big data characteristics, and the performance evaluation criterion is based only on price/load data, which is not large. With the consideration of big data characteristics, the classification accuracy needs to be further improved [3], [10].

A. *MOTIVATION*

This work investigates the ETD issue in the smart grid. The main objective is to identify fraudulent and honest consumers using big data from the smart grid. An SVM underpinned framework is proposed to solve the challenging binary classification task efficiently. To divide the given data into correct classes (honest and fraudulent), SVM tries to find a hyperplane with support vector help. Although SVM is the best-suited approach for the binary classification problem, subsequent challenges need to be tackled to achieve the problem's higher accuracy.

- ***Computational overhead***: In a work, Hu et al. [11] investigated that SVM performance is adversely affected by unreliable information and burdens the model with computational overhead. In the ETD problem, extraneous and redundant features increase computational overhead and make the classifier's training process difficult, decreasing the classification accuracy.
- ***Difficulty in handling sparse matrices***: In SVM, three super parameters, namely kernel parameter, intensive loss function, and cost penalty, control the classifier performance. To obtain optimum values, tuning these super parameters is a relatively tricky task for higher accuracy and better efficiency. Two well-known methods, namely cross-validation and gradient descent, are used to adjust SVM's super parameters [3]. However, both methods make the converging process hard and bring computational complexity.

B. *MAIN CONTRIBUTIONS*

To address the challenges mentioned above in ETD, we propose an integrated framework for electricity theft detection. As shown in Fig. 1, the two modules of the integrated framework are data preparations and final classification. First of all, data preparation performs interpolation, normalization, and balancing tasks. Precisely, data interpolation fills the missing values and brought consistency in the data set. Afterward, the data normalization (puts the values between 0–1) is performed to bring uniformity. Once the data preparation step is completed, the processed data is sent to the classifier. We chose SVM because it performs well on the classification task. SVM is very sensitive to the value of the hyperparameter. For this purpose, we employ a Bayesian Optimization algorithm-we name Bayesian SVM (BSVM) to detect electricity theft accurately. Our main contributions for higher accuracy and computational efficiency are listed below:

1) To achieve higher accuracy, an integrated framework based on two modules is proposed. Due to the cascading effect, smart meter theft data is efficiently handled and analyzed.
2) To achieve this, we first perform the data preparation step, which consists of filling missing values, data standardization, and handling imbalance class.
3) To improve standard SVM performance, a Bayesian Optimization algorithm is used to tune the hyperparameters. The BSVM has higher accuracy and computational efficiency than the basic SVM and recent machine learning techniques in the proposed area.
4) For performance evaluation, extensive simulations on real-world data traces of smart meter data have been considered. The numerical results show that the proposed model shows better performance statistics than benchmark approaches.

C. *ORGANIZATION*

Inspired by [3], Fig. 1 shows the framework of the proposed system based on two modules: data preparation and classification. Section II describes the data preparation module of the proposed ETD framework. Similarly, Section III demonstrates the SVM classifier and its enhancement with Bayesian Optimizer. The proposed framework for ETD is verified with multiple scenarios in Section IV. Finally, Section V concludes this work.

II. DATA PREPARATIONS

The preliminary analysis of data is a mandatory step in high dynamic time series analysis, which includes: imputation, data standardization, handling imbalance claas data. The details of these methods is given below.

1) ***Handling Missing Values****:* The electricity consumption record of consumers is usually composed of incomplete information or missing values. The reasons behind the issue may be the failure of hardware and corruption of data. In high time-series data, the missing values can not be dropped; however, the imputation is performed synthetically to fill these values. In most cases, the filling of missing values is performed through averaging. In this paper, the missing values are recovered through interpolation method [12] calculated below,

$$f(x_i) = \begin{cases} \left(\frac{x_{i-1}+x_{i+1}}{2}\right) & if\ x_i \in NaN, x_{i-1}, x_{i+1} \notin Nan \\ x_i & otherwise, \end{cases} \tag{1}$$

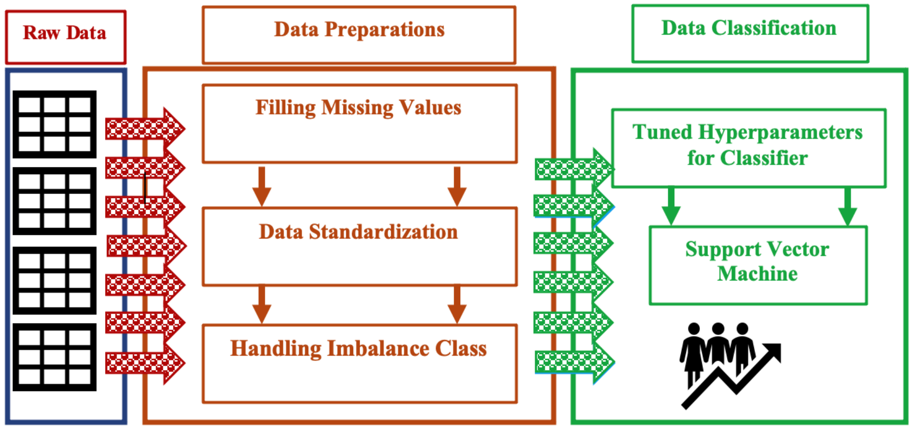

Fig. 1: **Proposed Framework for Electricity Theft Detection**

where, x_i is the recorded or missed (null) observation in the dataset. The null value is represented as NaN. If z_i is null then it is filled according to Eq. 1.

*2) **Handling Outliers**:* In the State Grid Corporation of China (SGCC) dataset, there are numerous outliers due to which data is skewed, and hence training process becomes complex. To avoid overfitting issue, these outliers must be identified and removed while preparing data for training. The "three-sigma rule of thumb" proposed in [13] is utilized for detecting and recovering the outliers. This is expressed as follows,

$$x_{out} = \begin{cases} X & if\ x_i > X \\ x_i & otherwise, \end{cases} \tag{2}$$

where, X is $Avg(x_i + 2\sigma(x_i))$ in above equation.

*3) **Data Standardization**:* Similarly, the data standardization is performed by min-max normalization method [13] as follows,

$$x_{new} = \frac{x_i - min(x)}{max(x) - min(x)} \tag{3}$$

*4) **Handling Imbalance Data**:* One of the critical problems in smart meter data is the majority class's domination (honest consumers) compared to the minority class (thieves). In such a scenario, the distribution is not normal and skewed towards the majority class because of an unclear decision boundary [14]. The classifier would become biased, may not learn critical features, and tend to become overfit. Traditional methods to deal with such issues are random under-sampling and random oversampling. However, these methods are not preferred because of specific problems, namely computational overhead, under and overfitting. Considering the nature of the problem, we opt for a relatively new class balancer approach that combines the properties of SMOTE and Tomek Links techniques; we name the new technique STLU. This technique has not been utilized in ETD strategies for class balancing to the best of our knowledge. In STLU, SMOTE is an oversampling technique which synthesizes new plausible examples in the majority class. In contrast, Tomek Links identifies different nearest neighbors' classes in a dataset and removes majority class samples to achieve a suitable balance [15].

III. CLASSIFICATION

This module describes the final classification task via the processed data. We chose SVM because it is one of the most adopted, robust and efficient machine learning methods to provide a higher aclassification accuracy. We assume a matrix of electricity consumption data as follows,

$$\mathrm{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & 22 & \dots & x_{2n} \\ & & & \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \tag{4}$$

where, rows and columns represent the time stamps and the feature index of the data, respectively. Hence, the predicted component of ith day of data's jth component can be represented as x_{ij}. The matrix can also be formulated as,

$$\mathrm{X} = \begin{bmatrix} \overrightarrow{t_1} \\ \overrightarrow{t_2} \\ . \\ . \\ . \\ \overrightarrow{t_m} \end{bmatrix} \tag{5}$$

where,

$$\overrightarrow{t_k} = [x_{k1}, x_{k2}, ...x_{kn}]\ k \in [1, m]. \tag{6}$$

For a given training set $\{(x_i, y_i)\}_{i=1}^{N}$ (x_i and y_i represents samples and target classes) with binary output $y_i = \pm 1$. The classification problem is investigated in the following eqautions [3],

$$f(x, c) = \sum_{i=1}^{N} c_i \lambda_i(x) + b, \tag{7}$$

where, b depends on data distribution and $c_i^{\infty}(i=1,2,...)$ are classifier parameters to be adjusted. Eq. 7 defines a hyperplan in N dimensional space. The regularized risk function is calculated using following equation,

$$w(c)=\frac{\sum_{i=1}^{N}|y_i-f(xi,c)|_{\in}+\mu c^2}{N} \tag{8}$$

where, $\in$ represent parameter for intensive loss function, μ is a constant, and y_i is the actual class. To obtain parameter c, reguarized risk function minimzation is required for which the robust error function is calculated below,

$$x=\begin{cases} 0 & if|y_i-f(xi,c)<\in \\ |y_i-f(xi,c)| & otherwise \end{cases} \tag{9}$$

A. *OPTIMAL CLASSIFICATION with BAYESIAN OPTIMIZATION*

In SVM, we aim to minimize the regularized risk function. The regularized risk function has a strong relationship with three super parameters: the type of SVM kernel parameter (σ), Cost penalty (c) and the Intensive loss function ($\in$). The need for parameter optimization is undeniable, and computational efficiency is achievable if optimal values for these super parameters are chosen. In the past, various methods such as Cross-Validation (CV), Grid Search (GS), Gradient Descent (GD), and Heuristic Algorithms are proposed to adjust super parameters. However, these methods may cause a problematic convergence process due to high computational overhead. Also, CV, GS, and RS methods are comparatively ineffective because of random search and not updated on the previous best value to choose the next hyperparameters selection. For this purpose, a reliable Bayesian Optimization algorithm is chosen to tune SVM's super parameters.

The Bayesian approach is chosen for parameter optimization because it is more directed, faster, and predict according to the posterior. Bayesian optimization improves the hyper-parameter selection by making use of earlier experiments. First, it constructs a probabilistic model of the function with super parameters and evaluates it on the validation test. With multiple iterations, the Bayesian Optimizer gathers relevant information about the optimal locations, with a perfect balance between exploration (super parameters likely to give uncertain outcome) and exploitation (expected optimum parameters). It provides better results in fewer iterations compared to the random and grid search algorithms. It starts with taking a history of super parameters settings $\lambda_n = \lambda_1, \lambda_2, \dots \lambda_n$ and respective function evaluation $y_1 = y_1, y_2, \dots y_n$ to acquire a new set of super parameters λ_{n+1}. In next iteration, λ_{n+1} is used as new population for model evaluation to get new function value y_{n+1}. Both the function values and super parameter values are saved to the history for execution of next iteration λ_{n+1}, y_{n+1}. In this way, the objective function's optimized value is achieved with a history of function evaluation and super parameters values.

With efficient data pre-processing and enhanced classification methodologies, th proposed framework is capable of performing ETD accurately. In the next section, a detailed analysis is given of real-world electricity theft data.

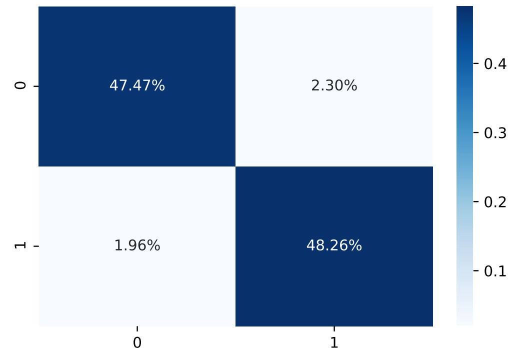

Fig. 2: **Confusion Matrix**

IV. NUMERICAL ANALYSIS

A. *SIMULATION SETUP*

This section investigates the proposed framework's performance, and the Python simulator is developed according to Section II framework. The simulation results are obtained on a platform with MAC i7, 16GB RAM, and 128GB hard disk. For this framework, input data is acquired from the most extensive power providing company in China, i.e., SGCC, from 2014 to 2016. With a daily electricity consumption profile of 42372 consumers, the record consists of over 38757 fair and the remaining 3615 fraudsters consumers.

B. *PERFORMANCE MATRIX*

The performance metrics are determined from the confusion matrix (CM), i.e., a matrix that describes different results in classification problems shown in Fig. 2. In a binary classification problem, the CM has four possible outcomes with two rows and two columns. These are true positive (TP), true negative (TN), false positive (FP), and false-negative (FN). TN and TP score mean that honest and dishonest consumers are identified accurately by the classifier. Similarly, FP and FN score means that the number of honest and dishonest consumers is misclassified. Based on CM results, the following equations calculate the performance of any classifier.

$$Precision=\frac{T^+}{T^+ + F^+} \tag{10}$$

$$Recall=\frac{T^+}{T^+ + F^-} \tag{11}$$

$$F_1\ Score = 2\times\frac{Precision\times Recall}{Precision + Recall} \tag{12}$$

$$Accuracy=\frac{T^+ + T^-}{T^+T^- + F^+ + F^-} \tag{13}$$

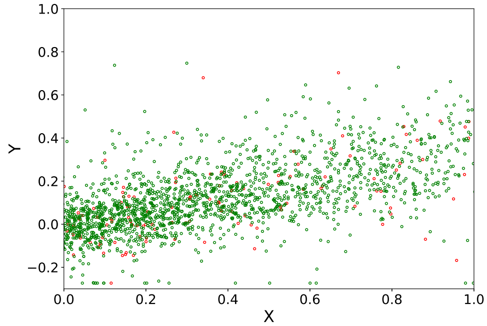

Fig. 3: **Imbalnce Data Before Sampling**

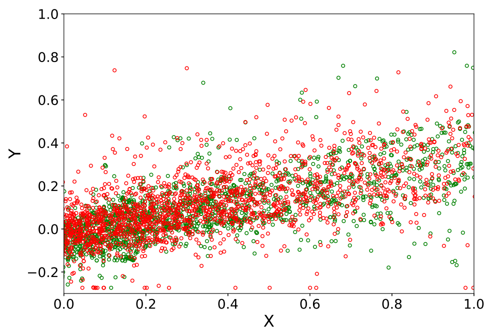

Fig. 4: **Balanced Data After Sampling**

C. *SIMULATION RESULTS*

1) IMPACT of HANDELING IMBALANCE CLASS: In an imbalanced class problem, one class significantly dominates the other category; hence, it results in the suppression of the minority class. Fig. 3 shows the difference between minority and majority classes before handling imbalance class. Clearly, the majority class (orange) customers are in a much higher ratio, and biased classification is expected. Without addressing

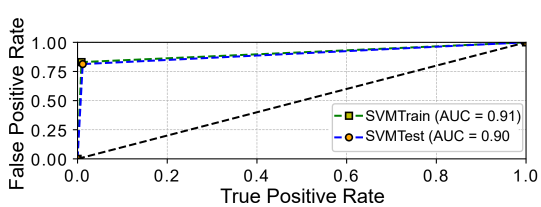

Fig. 5: **ROC-AUC Curve of SVM**

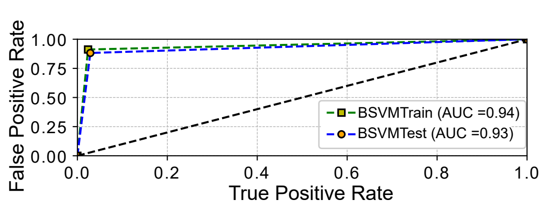

Fig. 6: **ROC-AUC Curve of BSVM**

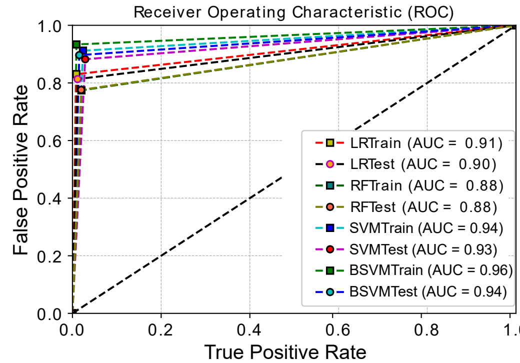

Fig. 7: **ROC-AUC-based Performance Comparison**

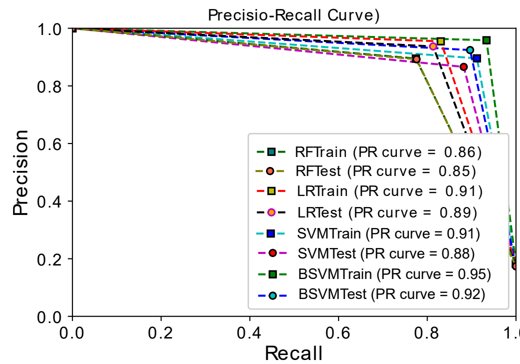

Fig. 8: **Precison-Recall Curve for ALL Methods**

the imbalance class issue, the value of AUC is 0.5850, for Precision is 0.7021 and that for Recall is 0.4453. The model fails to provide promising results, mainly while calculating Recall, when many fraud instances are misclassified as fair. To solve this problem, we apply STLU, which efficiently balances minority and majority class, and its impact is shown in Fig. 4. With balances data, model training, and generalization improve, as shown in Table 1.

2) The BSVM Performance of Theft Detection: SVM is a prevalent technique for classification tasks, and like any other machine learning method, its performance is mainly based on hyperparameters value. We tuned SVM's super parameters with Bayesian Optimizer, and the objective is to find an optimal hyperplane to distinguish different classes.

The Receiver Operating Characteristic (ROC) curve is the best performance metric used for detecting suspects in imbalanced class distribution [12]. It is the graphical representation of T+ rate and F+ rate and area under the ROC curve is called Area under the Curve (AUC). It separates the distribution of fraudulent class from fair class and expressed as follows,

$$AUC = \frac{\Sigma_i \in SR_i - \frac{1}{2}|S|(|S|+1)}{|S| \times |H|} \tag{14}$$

where, R_i denotes the rank of suspicion degree of fraud consumers in ascending order, $|S|$ and $|H|$ are the cardinality of suspicious and honest consumers.

The limits of ROC curve ranges from 0 to 1. The ideal situation arises, when no curve overlaps each other. AUC

TABLE I: Comparision among BSVM and Other Benchmark Schems

Methods	Training Ration 60 %					Training Ratio 80 %				
	Precision	Recall	F1	Accuracy	AUC	Precision	Recall	F1	Accuracy	AUC
Logistic Regression	0.713	0.710	0.688	0.700	0.700	0.770	0.725	0.725	0.770	0.720
Random Forest	0.688	0.677	0.687	0.687	0.755	0.751	0.753	0.747	0.757	0.755
Support Vector Machine	0.680	0.689	0.683	0.682	0.690	0.680	0.689	0.683	0.682	0.690
Bayesian Support Vector Machine	0.969	0.915	0.941	0.941	0.938	0.969	0.915	0.941	0.941	0.938

approaches 1, demonstrates the validity of classifier, while AUC less than 0.5 shows that the classifier does not have the ability to discriminate the classes. Figs. 5 and 6 show ROC-AUC curves of SVM and BSVM. The AUC of BSVM has been significantly improved both for training and testing. The Bayesian Optimizer optimizes SVM's super parameters jointly. Therefore, BSVM performance is better both in training and testing. Simultaneously, the AUC of SVM for training and testing are 0.91 and 0.90, whereas, for BSVM, these values are 0.94 and 0.93, respectively. This demonstrates that the acquired results are improved if Bayesian Optimization is used to find the SVM classifier's hyperparameters' values.

*3) **BSVM Performance Comparision with Benchmark Schemes**:* We compare the performance of BSVM and standard SVM with two benchmark classifiers, namely RF and LR. Figs. 7 and 8 compare ROC-AUC and Precision-Recall curves for all techniques. In this case, BSVM achieves higher accuracy for training and testing, which is up to 0.95 and 0.92, respectively. Furthermore, Fig. 2 shows that the ratio of FPR is only 2.30%, which is significantly less and acceptable for real-world scenarios. It implies that the proposed approach is reliable enough to identify fraudulent consumers in the electrical network.

*4) **BSVM Performance on Different Train/Test Data Sets**:* Table I provides an overview of each classifier's performance with an increasing training ratio of 60% and 80%. All obtained results of traditional classifiers such as LR, RF, SVM an expanding trend. Investigating the results, it is observed that the increase in training instances enhances traditional classifiers' performance. Moreover, it is clear in Table I that the proposed model is successfully applied to small-sized datasets and immensely large-sized datasets.

V. CONCLUSION

An accurate and reliable ETD method is essential for the electric power industry's planning and decision-making process. In this work, the smart grid's electricity theft detection problem is investigated via the combined effect of feature pre-processing and improved classification modules. Precisely, the smart meter data missing and inconsistent values are adjusted with data interpolation and standardization techniques. Additionally, the class imbalance problem is resolved with a newly developed combined over and under sampling technique. Finally, the Naive Optimizer obtains suitable values for cost penalty, kernel function, and intensive loss function automatically and efficiently for SVM. The numerical results show that our proposed framework is more accurate than LR, DT, and SVM and can perfectly be applied to industrial applications.

Acknowledgements: The authors acknowledge funding support from COMSATS University Islamabad, Pakistan and Lancaster University, UK to support this project.

REFERENCES

[1] Khan, Inam Ullah, et al. "Big Data Analytics Based Short Term Load Forecasting Model for Residential Buildings in Smart Grids." IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2020.
[2] Mohammadi, Mohsen, et al. "Small-scale building load forecast based on hybrid forecast engine. " Neural Processing Letters 48.1 (2018): 329-351.
[3] Khan, Inam Ullah, et al. "Heuristic Algorithm Based Optimal Power Flow Model Incorporating Stochastic Renewable Energy Sources."IEEE Access 8 (2020): 148622-148643.
[4] Gao, Wei, et al. "Different states of multi-block based forecast engine for price and load prediction." International Journal of Electrical Power & Energy Systems 104 (2019): 423-435.
[5] Wang, Kun, et al. "Robust big data analytics for electricity price forecasting in the smart grid." IEEE Transactions on Big Data 5.1 (2017): 34-45.
[6] Buzau, Madalina-Mihaela, et al. "Hybrid deep neural networks for detection of non-technical losses in electricity smart meters." IEEE Transactions on Power Systems 35.2 (2019): 1254-1263.
[7] Buzau, Madalina Mihaela, et al. "Detection of non-technical losses using smart meter data and supervised learning."IEEE Transactions on Smart Grid 10.3 (2018): 2661-2670.
[8] Li, Shuan, et al. "Electricity Theft Detection in Power Grids with Deep Learning and Random Forests." Journal of Electrical and Computer Engineering 2019 (2019).
[9] Buzau, Madalina Mihaela, et al. "Detection of non-technical losses using smart meter data and supervised learning." IEEE Transactions on Smart Grid 10.3 (2018): 2661-2670.
[10] Khan, Zahoor Ali, et al. "Electricity Theft Detection Using Supervised Learning Techniques on Smart Meter Data." Sustainability 12.19 (2020): 8023.
[11] Han, Hu, Jianwu Dang, and Enen Ren. "Comparative study of two uncertain support vector machines." 2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI). IEEE, 2012.
[12] Zheng, Zibin, et al. "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids." IEEE Transactions on Industrial Informatics 14.4 (2017): 1606-1615.
[13] Nam, Hyeonseob, and Hyo-Eun Kim. "Batch-instance normalization for adaptively style-invariant neural networks." Advances in Neural Information Processing Systems. 2018.
[14] J. Ma and X. Ma, "A Review of Forecasting Algorithms and Energy Management Strategies for Microgrids" Systems Science and Control Engineering, Vol. 6, No. 1, 2018, pp. 237-248.
[15] Machine Learning Mastery. Accessed: Dec. 1, 2020. Available: https://machinelearningmastery.com/combine-oversampling-and-undersampling-for-imbalanced-classification/